2

MRC Technical Summary Report #2327

USING EXTERNAL INFORMATION IN LINEAR REGRESSION, WITH A COMMENTARY ON RIDGE REGRESSION I. MIXED ESTIMATION AND BAYESIAN METHODS

Toby J. Mitchell and Norman R. Draper

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

January 1982

Received July 18, 1980

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

USING EXTERNAL INFORMATION IN LINEAR REGRESSION, WITH A COMMENTARY ON RIDGE REGRESSION
I. MIXED ESTIMATION AND BAYESIAN METHODS

Toby J. Mitchell* and Norman R. Draper

Technical Summary Report #2327
January 1982

ABSTRACT

In this article, we discuss ways of using "dummy data" and mixed estimation (Theil and Goldberger, 1961) to bring external information formally into linear regression problems when the experimental data/model are inadequate. This is a useful way of attacking the same practical problems that motivated the development of ridge regression (Hoerl and Kennard, 1970).

The main practical issues considered are (i) what form should the "dummy data" take?, and (ii) how much weight should it be given relative to the experimental data? When specific prior information is unavailable, it is suggested that the dummy data should reflect a preference for "stable" response functions and it is shown how this can be accomplished. Guidelines for the choice of the weighting parameter k (equivalent to the choice of the ridge parameter in ridge regression) are given. Upper limits for k are based on various tests of compatibility between the external (dummy) data and the experimental data. Lower limits for k are determined by the inadequacy of the data/model for the purpose(s) of the analysis. Finally, the parallel Bayesian approach is discussed, with emphasis on Box's (1980) framework of model estimation and criticism.

AMS(MOS) Subject Classification: 62J99

Key Words: Collinearity; Dummy data; External information in regression; Ridge regression.

Work Unit Number 4 - Statistics and Probability

*Oak Ridge National Laboratories, P. O. Box Y, Oak Ridge, TN 37830.

Sponsored by the United States Army under Contract Nos. DAAG29-80-C-0041 and DAAG29-80-C-0113.

SIGNIFICANCE AND EXPLANATION

We discuss ways of using "dummy data" to bring external information formally into linear regression problems when the experimental data/model are inadequate. The main practical issues are (i) what form should the "dummy data" take?, and (ii) how much weight should it be given relative to the experimental data? When specific prior information is unavailable, it is suggested that the dummy data should reflect a preference for so-called "stable" response functions, and it is shown how this can be accomplished. Guidelines for the choice of weighting are also given. Upper limits for a weighting parameter are based on various tests of compatibility between the external (dummy) data and the experimental data. Lower limits are determined by the limits of inadequacy of the data/model for the purpose(s) of the analysis. The parallel Bayesian approach is also discussed.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

USING EXTERNAL INFORMATION IN LINEAR REGRESSION, WITH A COMMENTARY ON RIDGE REGRESSION

I. MIXED ESTIMATION AND BAYESIAN METHODS

Toby J. Mitchell* and Norman R. Draper

1. INTRODUCTION.

In this report we focus on the problems that arise in linear regression when there is insufficient information in the data/model to obtain useful estimates of the regression coefficients and/or linear functions of them. Methods for augmenting the experimental data with external "information" lead to a form of ridge regression which differs in several major respects from that which is currently practiced.

We shall consider models for the observed response variable y_d, given associated "predictor data" d, to be of the form:

$$y_d = \eta_d + \varepsilon_d, \qquad (1.1)$$

where

$$\eta_d = \sum_{i=1}^{p} \beta_i \, x_i(d) = \underset{\sim}{x}'(d)\underset{\sim}{\beta}, \qquad (1.2)$$

the p <u>predictor variables</u> x_i are specified functions of d , the <u>regression coefficients</u> $\{\beta_i\}$ are unknown constants, and ε_d is a random variable with mean 0 and variance σ^2. For $d \neq d'$, ε_d is independent of $\varepsilon_{d'}$. The error variance σ^2, assumed to be constant for all d , and is generally unknown.

We assume that (1.1) and (1.2) hold over a <u>model region</u> R in the space of the predictor data. We shall refer to (1.2) as the <u>model function</u>, although it is really a family of model functions, indexed by the parameter vector $\underset{\sim}{\beta}$. We shall view the problem of "estimating" $\underset{\sim}{\beta}$ to be the problem of choosing from this family a single model function which is to be used to make inferences about η_d given $d \in R$, or more generally, to make inferences about specified linear functions of the β's .

*Oak Ridge National Laboratories, P. O. Box Y, Oak Ridge, TN 37830.

Sponsored by the United States Army under Contracts Nos. DAAG29-80-C-0041 and DAAG29-80-C-0113.

Given a sample of n independent observations $\underset{\sim}{y}' = (y_1, y_2, \cdots, y_n)$, where y_u is associated with known d_u, $u = 1,2, \cdots, n$, (1.1) and (1.2) imply the statistical model for the experimental data:

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} \qquad (1.3)$$

$$V(\underset{\sim}{y}) = \sigma^2 \underset{\sim}{I} \;, \qquad (1.4)$$

where the matrix $\underset{\sim}{X}$ is $n \times p$ with (u,i)th element X_{ui} equal to $x_i(d_u)$ and $\underset{\sim}{I}$ is the $n \times n$ identity matrix.

The most common estimation procedure follows from the least squares criterion: choose $\underset{\sim}{\beta}$ to minimize the sum of squares: $SS(\underset{\sim}{\beta}) = \|\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}\|^2$. This yields the least squares estimator

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y} \;, \qquad (1.5)$$

where we assume that $\underset{\sim}{X}'\underset{\sim}{X}$ is non-singular so that $\hat{\underset{\sim}{\beta}}$ is unique. (If $\underset{\sim}{X}'\underset{\sim}{X}$ were singular, (1.5) would contain a generalized inverse, and $\hat{\underset{\sim}{\beta}}$ would not be unique.)

The least-squares estimator may be justified on several statistical grounds, the most prominent of which are:

(i) The Gauss-Markov Theorem: Among all unbiased estimators for $\mu = \underset{\sim}{t}'\underset{\sim}{\beta}$ that are linear in the y's, $\underset{\sim}{t}'\hat{\underset{\sim}{\beta}}$ has minimum variance, no matter what $\underset{\sim}{t}$ is. (An alternative version that does not impose unbiasedness is: Among all linear estimators for μ, $\underset{\sim}{t}'\hat{\underset{\sim}{\beta}}$ minimizes the maximum mean squared error of estimation (Barnard, 1977).) Since the purpose of the model is often to estimate linear functions of the β's, this is generally considered a strong argument in favor of least squares, although the restriction to linear estimators seems arbitrary.

(ii) If $\underset{\sim}{y}$ has a multinormal distribution with mean and variance-covariance matrix given by (1.3) and (1.4), then $\hat{\underset{\sim}{\beta}}$ is the maximum likelihood estimator of $\underset{\sim}{\beta}$. Intuitively, we would expect the model function that is favored most by the available data, i.e., the maximum likelihood model, to serve at least as well as any other in making inferences about general linear functions of the β's.

These arguments in favor of least squares lose force if the model as given by (1.1) and (1.2) is inadequate, or if, in the case of (ii), $\underset{\sim}{y}$ is not normally distributed. Minimum bias estimation (Karson, Manson, and Hader, 1969; Kupper and Meydrech, 1974) and robust regression (Huber, 1981) are alternatives to least squares in such situations. We shall not consider these methods here, since we shall assume that the model assumptions (including normality, when needed) are satisfied. Even then, least squares estimation may seem unsatisfactory because the variances of the $\hat{\beta}$'s may be extremely high for certain configurations of the d's. Various alternative estimation methods have been proposed for these situations also; predominant among them are variable selection and ridge regression. (See Hocking (1976) for an excellent overview of these methods.)

The motivation for our work here is our dissatisfaction with the way ridge regression is perceived and used. There is an enormous literature on this subject, stemming from the basic papers of Hoerl and Kennard (1970a, 1970b). Unfortunately, much of this work presents ridge regression primarily as a biased estimation technique, based on the experimental data alone, that is "superior" in some sense to least squares estimation. This approach has misled users and authors alike. The main problem in situations where ridge regression is often invoked is one of lack of information in the data (with respect to the choice of a single model from the given family), and <u>not</u> with least squares estimation per se. External information or additional assumptions are needed; what form should these take and how should they be implemented? These are the topics of Part I of this report. We have deliberately written Part I almost as if ridge regression did not exist, in order to set out first the fundamentals that underlie our point of view. However, we do consider only the kind of external information that leads to "ridge-type" estimators. In Part II, we shall offer a commentary on some specific aspects of ridge regression, adding further support to recent criticisms of various ridge regression myths (Thisted 1976; Draper and Von Nostrand, 1979; Smith and Campbell, 1980; Smith, 1980; Egerton and Laycock, 1981).

Certainly the idea of considering ridge regression as a means of bringing external information into the estimation procedure is not new. Hoerl and Kennard (1970a) mentioned a Bayesian derivation of ridge estimation, and Lindley and Smith (1972) have provided an extensive Bayesian treatment. We shall rely more on the "mixed estimation" approach of Theil and Goldberger (1961), in that the external information is introduced by means of dummy data rather than prior distributions. We offer (in Section 3) a rationale for specifyin such data in the absence of specific external information about the regression coefficients. In Section 4, we suggest some guidelines for selecting the amount of influence, or weight, to give the dummy data; this is equivalent to the choice of the ridge parameter k in ridge regression. Bayesian methods are discussed in Section 5 with emphasis on Box's (1980) illuminating approach to model estimation and criticism.

2. BRINGING IN EXTERNAL INFORMATION

2.1. The need for external information.

Suppose we wish to estimate one or more linear functions of the regression coefficients. We shall denote these by $\mu_1, \mu_2, \cdots, \mu_m$, where $\mu_i = \underset{\sim}{t}_i'\underset{\sim}{\beta}$ and the t's are known. If the response vector $\underset{\sim}{y}$ is normally distributed:

$$\underset{\sim}{y} \sim N(\underset{\sim}{X}\underset{\sim}{\beta}, \underset{\sim}{I}\sigma^2), \tag{2.1}$$

the set of "acceptable" values of $\underset{\sim}{\beta}$ lies inside the ellipsoid:

$$(\underset{\sim}{\beta} - \hat{\underset{\sim}{\beta}})'\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{\beta} - \hat{\underset{\sim}{\beta}}) \leq a \tag{2.2}$$

for some suitably chosen positive constant a. This follows from a standard likelihood argument which requires only that every "acceptable" $\underset{\sim}{\beta}$ has a likelihood greater than that of every "unacceptable" $\underset{\sim}{\beta}$. By maximizing and minimizing $\underset{\sim}{t}_i'\underset{\sim}{\beta}$ subject to (2.2), we find that the "acceptable" values for μ_i must be in the interval with end points:

$$\underset{\sim}{t}_i'\underset{\sim}{\beta} \pm (a\ \underset{\sim}{t}_i'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{t}_i)^{1/2}. \tag{2.3}$$

The choice $a = \sigma^2\chi^2_{1;\alpha}$ in (2.3) where $\chi^2_{1;\alpha}$ is the upper α% point of the chi-squared distribution with one degree of freedom, results in the usual $100(1-\alpha)$% confidence interval for μ_i when σ^2 is known. The other factor that determines the interval width, namely,

$$w_i = \underset{\sim}{t}_i'(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{t}_i \tag{2.4}$$

is affected jointly by the model and "design", (i.e., the configuration of the d's), which express themselves through $\underset{\sim}{X}$.

If w_i in (2.4) is such that the interval (2.3) is too wide to be of value, we conclude that the data and model assumptions alone are inadequate for the purpose of estimating $\mu_i = \underset{\sim}{t}_i'\underset{\sim}{\beta}$. Ideally, we would like to have more experimental data, but we shall assume here that this is not possible and that our only recourse is to bring in external, often vague, information about $\underset{\sim}{\beta}$. For another purpose (i.e., another $\underset{\sim}{t}_i$), of course, there might be no need at all for additional information or assumptions.

We have deliberately avoided presenting w_i as the variance of $\underset{\sim}{t}'\hat{\underset{\sim}{\beta}}$ (divided by σ^2) to emphasize that its role as an indicator of the information provided by the data/model with respect to μ_i is independent of whatever point estimation procedure for $\underset{\sim}{\beta}$ is used. If w_i is "too wide", the problem cannot be alleviated by blaming least squares estimation and seeking a "better" estimator in some class of biased estimators. External information must be used.

2.2. Mixed estimation.

Our discussion will center on the use of "mixed estimation", proposed by Theil and Goldberger (1961) and developed further by Theil (1963), as a vehicle for bringing in external information. (We shall briefly summarize the Bayesian parallel in Section 5.)

For mixed estimation, one augments the observed data with an n_0-vector of "dummy data" $\mathbf{y}_0$, i.e., guessed values of linear functions of the β's. These guesses are modeled as:

$$\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta} + \mathbf{e}_0 \tag{2.5}$$

where $\mathbf{X}_0$ is a specified $n_0 \times p$ matrix and

$$\mathbf{e}_0 \sim N(\mathbf{0},\sigma_0^2\mathbf{V}_0). \tag{2.6}$$

We shall consider $\mathbf{V}_0$ to be known throughout; its choice will be considered in Section 3. The guesses are consistent in that there exists some $\boldsymbol{\beta}^*$ (not necessarily the "true" $\boldsymbol{\beta}$) such that

$$\mathbf{X}_0\boldsymbol{\beta}^* = \mathbf{y}_0. \tag{2.7}$$

For known σ^2 amd σ_0^2, the mixed estimator is the weighted least squares estimator of $\boldsymbol{\beta}$ based on all the data:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{W}+\mathbf{W}_0)^{-1}(\mathbf{W}\hat{\boldsymbol{\beta}} + \mathbf{W}_0\boldsymbol{\beta}^*) \tag{2.8}$$

where

$$\mathbf{W} = \mathbf{X}'\mathbf{X}/\sigma^2 \tag{2.9}$$

and

$$W_0 = \mathbf{X}_0'\mathbf{V}_0^{-1}\mathbf{X}_0/\sigma_0^2. \tag{2.10}$$

If we let

$$k = \sigma^2/\sigma_0^2 \tag{2.11}$$

and

$$\mathbf{T} = \mathbf{X}_0'\mathbf{V}_0^{-1}\mathbf{X}_0. \tag{2.12}$$

Then (2.8) can be written

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + k\mathbf{T})^{-1}(\mathbf{X}'\mathbf{y} + k\mathbf{T}\boldsymbol{\beta}^{*}). \qquad (2.13)$$

The variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ is

$$V(\tilde{\boldsymbol{\beta}}) = (\mathbf{W} + \mathbf{W}_0)^{-1} = \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{T})^{-1}. \qquad (2.14)$$

When σ^2 is not known, Theil and Goldberger (1961) and Theil (1963) proposed to replace σ^2 by s^2, the residual mean square from the least squares analysis of the experimental data. They also mentioned an iterative approach in which a new estimate of σ^2 and hence a new weight matrix $\mathbf{W}$ is derived from the residual mean square after each iteration. Under this procedure, successive estimates of $\tilde{\beta}$ converge to the maximum likelihood estimate of $\boldsymbol{\beta}$ based on all the data. Unfortunately, this estimate of $\boldsymbol{\beta}$ is not linear in the observations, so further statistical results are difficult to obtain.

We shall adopt here a more tractable approach, and treat σ_0^2 as an unknown multiple of σ^2:

$$\sigma_0^2 = k^{-1}\sigma^2. \qquad (2.15)$$

For fixed k, this will allow us to make inferences using standard weighted least squares procedures, where the experimental data are given weight 1 and the dummy data are given weight k. Thus $\tilde{\boldsymbol{\beta}}$ is given by (2.13) with variance-covariance matrix given by (2.14). The estimate of σ^2 is based on the weighted residual mean square

$$\tilde{\sigma}^2 = (\nu s^2 + q)/(\nu + n_0) \qquad (2.16)$$

where

$$q = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*})'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{T})^{-1}k\mathbf{T}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}), \qquad (2.17)$$

and $\nu = n-p$ is the number of degrees of freedom for s^2. Confidence intervals for linear functions of the β's can be constructed in the usual way, using Student's t distribution with $\nu+n_0$ degrees of freedom. Thus, a $100(1-\alpha)\%$ confidence interval for $\mu = \mathbf{t}'\boldsymbol{\beta}$ is given by

$$\mathbf{t}'\tilde{\boldsymbol{\beta}} \pm t_{\nu+n_0;\alpha/2}[\mathbf{t}'(\mathbf{X}'\mathbf{X} + k\mathbf{T})^{-1}\mathbf{t}]^{1/2}\tilde{\sigma}. \qquad (2.18)$$

We note that if $\mathbf{X}$ has been "centered and scaled" so that $\mathbf{X}'\mathbf{X}$ is in the form of a correlation marix, and if $\mathbf{T} = \mathbf{I}$ and $\boldsymbol{\beta}^{*} = \mathbf{0}$, then $\tilde{\boldsymbol{\beta}}$ reduces to the standard Hoerl-Kennard (1970a) ridge estimator. Proponents of classical ridge regression, while admitting that ridge regression *can* be viewed as the result of using external information, would deny the formal use of that information in determining confidence intervals. They would then arrive at the standard "least squares' confidence intervals rather than (2.18). (See Obenchain, 1977.)

The difficulties in finding a "best" value for k are well known to followers of ridge regression. Our approach here will be to regard k as *fixed by assumption,* then check that assumption for compatibility with the experimental data, much in the spirit of "model criticism" put forward by Box (1980). We shall discuss these procedures in Section 4, after we deal with the problem of specifying $\underset{\sim}{\Sigma}$ and $\underset{\sim}{\beta}$.

3. USE OF PRIOR PREFERENCES TO DETERMINE $\underset{\sim}{T}$ and $\underset{\sim}{\beta}^{*}$.

3.1. Prior preferences.

We shall assume that we can establish (by the approach to be suggested in Section 3.2) a convention for ranking proposed values of $\underset{\sim}{\beta}$ a priori, where the ranking is such that $\underset{\sim}{\beta}^{(1)}$ is preferred to $\underset{\sim}{\beta}^{(2)}$ if and only if

$$(\underset{\sim}{\beta}^{(1)}-\underset{\sim}{\beta}^{0})'\underset{\sim}{T}^{0}(\underset{\sim}{\beta}^{(1)}-\underset{\sim}{\beta}^{0}) < (\underset{\sim}{\beta}^{(2)}-\underset{\sim}{\beta}^{0})'\underset{\sim}{T}^{0}(\underset{\sim}{\beta}^{(2)}-\underset{\sim}{\beta}^{0}) \qquad (3.1)$$

for specified $\underset{\sim}{\beta}^{0}$ and non-negative definite $\underset{\sim}{T}^{0}$. To ensure that the likelihood for the dummy data reflects these preferences, we would choose $\underset{\sim}{\beta}^{*} = \underset{\sim}{\beta}^{0}$ and $\underset{\sim}{T} = \underset{\sim}{T}^{0}$ in our mixed estimation procedure.

A special case of this kind of rule is Hoerl and Kennard's (1970) preference for "short" parameter vectors, expressed in (3.1) as $\underset{\sim}{\beta}^{0} = \underset{\sim}{0}$, $\underset{\sim}{T}^{0} = \underset{\sim}{I}$, where $\underset{\sim}{X}'\underset{\sim}{X}$ is in correlation form. This particular preference has been criticized (rightly, we believe) because the length of $\underset{\sim}{\beta}$ depends on the parametrization of the model. (See Smith and Campbell (1980).).

Here we present a rationale for choosing $\underset{\sim}{T}^{0}$ and $\underset{\sim}{\beta}^{0}$ which is in the same spirit as the Hoerl-Kennard preference, and at the same time produces a family of estimators that is invariant under reparametrizations of the model. This rationale, which is based on the concept of "stability" of the expected response η_d, also extends naturally to quadratic and higher order models. In some circumstances, of course, one's prior information may permit the choice of $\underset{\sim}{T}^{0}$ and $\underset{\sim}{\beta}^{0}$ to be made directly.

3.2. "Stable response" preferences.

Suppose R_s is a selected subregion of the model region R. We define the <u>instability</u> Γ of the response function η_d over R_s to be the variance of η_d induced by a uniform probability distribution $\phi(d)$ over R_s. (This is just the squared deviation of η_d from its average, integrated over R_s. We assume, of course, that R_s and the functions $x_i(d)$ are sufficiently well-behaved for Γ to exist.) We emphasize that Γ is an inherent property of the model function η_d ; it has nothing to do with the data.

From (1.2), Γ can be written as a quadratic form:

$$\Gamma = \underset{\sim}{\beta}'\,\underset{\sim}{U}\,\underset{\sim}{\beta} \qquad (3.2)$$

where

$$\mathbf{U} = V_\phi(\mathbf{x}(d)) = E_\phi(\mathbf{x}(d) - \boldsymbol{\xi})(\mathbf{x}(d) - \boldsymbol{\xi})' = \mathbf{M} - \boldsymbol{\xi}\boldsymbol{\xi}' \tag{3.3}$$

where

$$\mathbf{M} = E_\phi(\mathbf{x}(d)\mathbf{x}'(d)) \tag{3.4}$$

and

$$\boldsymbol{\xi} = E_\phi(\mathbf{x}(d)). \tag{3.5}$$

In the absence of any other external information about $\boldsymbol{\beta}$, we might simply express a preference for those $\boldsymbol{\beta}$'s that give a more stable response η_d over R_s, i.e. we choose $\mathbf{T}^0 = \mathbf{u}$ and $\boldsymbol{\beta}^0 = \mathbf{0}$ in (3.1). This leads, via mixed estimation, to a "stable response" (SR) estimator of form (2.13) with $\mathbf{T} = \mathbf{U}$ and $\boldsymbol{\beta}^* = \mathbf{0}$, i.e.,

$$\tilde{\boldsymbol{\beta}}_{SR} = (\mathbf{X}'\mathbf{X} + k\mathbf{U})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.6}$$

Now suppose the model function (1.2) is reparametrized as follows:

$$\eta_d = \sum_{i=1}^{p} \omega_i f_i(d) = \mathbf{f}'(d)\boldsymbol{\omega} \tag{3.7}$$

where

$$\mathbf{f}'(d) = \mathbf{x}'(d)\mathbf{A}, \quad \boldsymbol{\omega} = \mathbf{A}^{-1}\boldsymbol{\beta}, \tag{3.8}$$

$\mathbf{A}$ being a nonsingular p×p matrix. Then the appropriate $\mathbf{U}$-matrix is:

$$\mathbf{U}_f = V_\phi(\mathbf{f}(d)) = \mathbf{A}'\mathbf{U}\mathbf{A} , \tag{3.9}$$

the predictor data matrix is $\mathbf{X}\mathbf{A}$, and the SR estimator for $\boldsymbol{\omega}$ is:

$$\begin{aligned} \tilde{\boldsymbol{\omega}}_{SR} &= (\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} + k\mathbf{A}'\mathbf{U}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{Y} \\ &= \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X} + k\mathbf{U})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{A}^{-1}\tilde{\boldsymbol{\beta}}_{SR} \ . \end{aligned} \tag{3.10}$$

Thus the SR estimator is invariant under reparameterizations of the model. This is what we would expect, since the instability of the model function depends only on η_d and R_s , neither of which is affected by reparameterization.

It is interesting to examine the influence of various choices of the "region of stability" R_s on $\mathbf{U}$, and hence on $\tilde{\boldsymbol{\beta}}_{SR}$. In the three cases that follow we suppose that there is a constant term in the model, so it will be convenient to change notation slightly by writing the model function (1.2) as

$$\eta_d = \beta_0 + \sum_{i=1}^{p} \beta_i x_i(d) = \beta_0 + \underset{\sim}{x}'(d)\underset{\sim}{\beta} \qquad (3.11)$$

and the expected response vector (1.3) as

$$E(\underset{\sim}{y}) = \underset{\sim}{1}\beta_0 + \underset{\sim}{X}\underset{\sim}{\beta}, \qquad (3.12)$$

where $\underset{\sim}{1}$ is a vector of 1's. Note that p is now 1 less than the total number of coefficients in the model.

Case 1. R_s consists of the data points $d_1, d_2, \cdots, d_n$ themselves.

Since $\phi(d) = 1/n$ at every d_u in the data set, (3.5) yields $\underset{\sim}{\xi}' = (1, \bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$, where $\bar{x}_i = \sum_{u=1}^{n} x_{ui}/n$. It then follows from (3.3) that

$$\underset{\sim}{U} = \begin{bmatrix} 0 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \frac{1}{n}\dot{\underset{\sim}{X}}'\dot{\underset{\sim}{X}} \end{bmatrix} \qquad (3.13)$$

where $\dot{\underset{\sim}{X}}$ is obtained from $\underset{\sim}{X}$ by "centering" the columns, i.e.,

$$\dot{\underset{\sim}{X}} = (\underset{\sim}{I} - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}')\underset{\sim}{X} . \qquad (3.14)$$

It can be shown that the corresponding SR estimator (3.6) is given by:

$$\begin{aligned} \tilde{\underset{\sim}{\beta}}_{SR} &= \frac{n}{n+k}\hat{\underset{\sim}{\beta}} \\ \tilde{\beta}_{0,SR} &= \bar{y} - \sum_{i=1}^{p} \tilde{\beta}_{i,SR}\bar{x}_i . \end{aligned} \qquad (3.15)$$

The effect of the external data here is to "shrink" the least squares estimator.

Case 2. "Factorial" R_s defined by the levels of the observed predictor variables - first order model.

We shall assume here that η_d is a first-order model function, i.e. the predictors $x_i(d)$ in (3.11) are not functionally related. We now define the region of stability R_s to correspond to the factorial "design" generated by the observed levels of the individual predictors. Let $\Omega_i = \{x_{1i}, x_{2i}, \cdots, x_{ni}\}$, i.e., Ω_i is the set of observed values of x_i, and let $\underset{\sim}{\Omega}$ be the matrix whose rows are elements of the lattice $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_p$. This lattice has n^p elements, some of which may be identical. For example, if $\underset{\sim}{X} = \begin{bmatrix} 1 & 2 \\ 1 & 3.5 \\ 2 & -3 \end{bmatrix}$ then $\underset{\sim}{\Omega} = \begin{bmatrix} 1 & 2 \\ 1 & 3.5 \\ 1 & -3 \\ 1 & 2 \\ 1 & 3.5 \\ 1 & -3 \\ 2 & 2 \\ 2 & 3.5 \\ 2 & -3 \end{bmatrix}$.

We take R_s to be the set of points $\{\omega_j\}$ corresponding to the rows of $\underset{\sim}{\Omega}$, with $\phi(\omega_j) = 1/n^P$. (Note, however, that under this definition R_s depends on the parametrization of the model, so the invariance result (3.10) will not hold.)

Under the above conditions, $\underset{\sim}{U}$ as defined by (3.3) - (3.5) is:

$$\underset{\sim}{U} = \frac{1}{n}\left[\begin{array}{c|c} 0 & \underset{\sim}{0}' \\ \hline \underset{\sim}{0} & \underset{\sim}{S} \end{array}\right] \qquad (3.16)$$

where $\underset{\sim}{S}$ is a diagonal matrix with ith diagonal element

$$S_i = \sum_u (X_{ui} - \bar{X}_i)^2 , \qquad i = 1,2,\cdots,P . \qquad (3.17)$$

To see this, note that $U_{i+1,i+1}$ is, by (3.3),the variance of x_i with respect to the distribution ϕ. Under ϕ, the probability at X_{ui} is just $1/n$, so the mean of x_i is $\bar{X}_i$ and the variance of x_i is S_j/n. Moreover, because of the independence of x_i and x_j for all pairs $i \neq j$, the off-diagonal terms of $\underset{\sim}{S}$ are 0. Finally the constancy of $x_0 = 1$ under ϕ results in the zero values in the first row and first column of $\underset{\sim}{U}$.

With $\underset{\sim}{U}$ given by (3.16), the SR estimator (3.6) is given by:

$$\begin{aligned} \tilde{\underset{\sim}{\beta}}_{SR} &= (\underset{\sim}{Z}'\underset{\sim}{Z} + k\underset{\sim}{I})^{-1}\underset{\sim}{Z}'\underset{\sim}{Y} \\ \tilde{\beta}_{0,SR} &= \bar{y} - \sum_{i=1}^{P} \tilde{\beta}_{i,SR}\,\bar{X}_i \end{aligned} \qquad (3.18)$$

where $\underset{\sim}{Z}$ is the "correlation form" of $\underset{\sim}{X}$, i.e.,

$$\underset{\sim}{Z} = \dot{\underset{\sim}{X}}\,\underset{\sim}{S}^{-1/2} \qquad (3.19)$$

Note that (3.18) is the same as the Hoerl-Kennard family of ridge estimators. We consider this derivation to be a curosity rather than a rationale for using the Hoerl-Kennard estimators. Generally, we would prefer to define R_s so that it depends neither on the observed data nor on the parametrization of the model.

Case 3. Rectangular R_s - first order model.

In most regression problems, the predictor data d take the form of a vector of r "basic" predictor variables, and the region of stability R_s can be taken to be some geometrically convenient region in R. In this example, we again assume a first-order model, so $p = r$, and we suppose that the predictor variables have been coded so that

$$-1 \leq x_j(d) \leq 1 \iff d \in R_s \; . \qquad (3.20)$$

Taking $\phi(d)$ to be uniform over R_s, we find

$$\underset{\sim}{U} = \begin{bmatrix} 0 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \frac{1}{3}\underset{\sim}{I} \end{bmatrix} \; . \qquad (3.21)$$

With $\underset{\sim}{U}$ given by (3.21), $\tilde{\underset{\sim}{\beta}}_{SR}$ is given by

$$\begin{aligned} \tilde{\underset{\sim}{\beta}}_{SR} &= (\dot{\underset{\sim}{X}}'\dot{\underset{\sim}{X}} + \frac{k}{3}\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \\ \tilde{\beta}_{0,SR} &= \bar{y} - \sum_{i=1}^{p} \tilde{\underset{\sim}{\beta}}_{i,SR}\, \bar{x}_i \; . \end{aligned} \qquad (3.22)$$

We note that if one were to use the Hoerl-Kennard formula, after centering, but not scaling $\underset{\sim}{X}$, one would arrive at this same family of estimators. ($\underset{\sim}{X}$ has already been effectively scaled by the coding (3.20).) As in Case 2 above, however, the application of the stable response preference to higher order models results in a more complicated form of $\underset{\sim}{U}$, and the analogy with the Hoerl-Kennard approach no longer holds.

In the above Cases 1-3, we have given some examples of the derivation of the stable response family of estimators for different choices of the region of stability R_s. In Cases 1 and 2, which lead to the more familiar "shrunken" estimator and Hoerl-Kennard estimator, the data define ϕ. We prefer the approach taken in Case 3, however, since we feel that the notion of response surface stability over some region should not depend on the distribution of the current data over that region.

We offer the "stable response" preference primarily as a route to follow in the absence of specific prior information about individual β's or about the expected response η_d at certain points d R. The notion of instability of the true response surface over a specified region seems to us to lend itself to physical interpretation much more readily than the notion of length of the vector of regression coefficients, even though the choice of region of stability is still rather arbitrary. The operational advantage of the SR method is that one can apply it without having to worry about parametrization or about the specific meaning of individual coefficients.

3.3. Setting up the dummy data.

External preferences of the form (3.1) can be transformed into dummy data for mixed estimation by choosing $\underset{\sim}{X}_0$, $\underset{\sim}{V}_0$, and $\underset{\sim}{y}_0$ to satisfy (2.12) and (2.7) with $\underset{\sim}{T} = \underset{\sim}{T}_0$ and $\underset{\sim}{\beta}^* = \underset{\sim}{\beta}^0$. The n_0 rows of $\underset{\sim}{X}_0$ could, for example, be the eigenvectors corresponding to the positive eigenvalues λ_i, $i = 1,2,\cdots,n_0$, of $\underset{\sim}{T}$; $\underset{\sim}{V}_0^{-1}$ would then be a diagonal matrix with those eigenvalues on the diagonal. Since the analysis depends only on $\underset{\sim}{T}$ and $\underset{\sim}{\beta}^*$, the only point in choosing $\underset{\sim}{X}_0$, $\underset{\sim}{V}_0$ and $\underset{\sim}{y}_0$ in this way is to take advantage of standard computer routines for doing weighted least squares. The weight for the ith dummy observation would be $k\lambda_i$; each experimental observation would have weight 1.

4. SOME GUIDELINES FOR CHOOSING k.

4.1. Omnibus compatibility test.

Throughout this section, it will be convenient to express $\tilde{\beta}$, $\tilde{\sigma}^2$, and q, defined respectively by (2.13), (2.16), and (2.17), as functions of k.

Although we doubt that a "best" procedure for choosing k exists, we can present some guidelines. For example, we shall certainly not want to choose k so large that the external (dummy) data is incompatible with the experimental data. Theil's (1963) test statistic for detecting incompatibility is, after adapting for fixed k rather than fixed $\sigma_0^2\boldsymbol{v}_0$ and expressing in terms of $\boldsymbol{\beta}$ and $\boldsymbol{T}$:

$$\psi(k) = q(k)/n_0 s^2, \tag{4.1}$$

where q(k) is given by (2.17). Under the assumed model for the experimental and dummy data, $\psi(k)$ is distributed as $F_{n_0,\nu}$ where n_0 is the rank of $\boldsymbol{T}$. Large values of $\psi(k)$ indicate incompatibility.

This test for compatibility is equivalent to the usual test of the hypothesis: $E(\boldsymbol{y}_0) = \boldsymbol{X}_0\boldsymbol{\beta}$ against the alternative: $E(\boldsymbol{y}_0) \neq \boldsymbol{X}_0\boldsymbol{\beta}$, a test frequently used in regression to evaluate a subset of the observations (in this case $\boldsymbol{y}_0$) jointly for the presence of outliers. (Gentleman and Wilk, 1975.)

We propose that $\psi(k)$ be used to set an upper limit k_α on k, chosen so that

$$\psi(k_\alpha) = F_{n_0,\nu;\alpha} \tag{4.2}$$

for some suitably chosen percentage point (e.g. $\alpha = .10$) of the $F_{n_0,\nu}$ distribution.

Remark 1. If $k \le k_\alpha$ (i.e., if the external information "passes" the compatibility test) then $\tilde{\beta}(k)$ is within the usual (100(1-α)% confidence region for $\boldsymbol{\beta}$ based on the experimental data alone, i.e.,

$$(\tilde{\boldsymbol{\beta}}(k)-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\tilde{\boldsymbol{\beta}}(k)-\hat{\boldsymbol{\beta}}) \le p\, s^2F_{p,\nu;\alpha}. \tag{4.3}$$

That is, $\tilde{\boldsymbol{\beta}}(k)$ is "α-acceptable" in the sense of McCabe (1978). This follows from the fact that, for $k \le k_\alpha$,

$$(\tilde{\boldsymbol{\beta}}(k)-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\tilde{\boldsymbol{\beta}}(k)-\hat{\boldsymbol{\beta}}) < q(k) \le n_0s^2F_{n_0,\nu;\alpha} \le ps^2F_{p,\nu;\alpha} \tag{4.4}$$

since it can be shown that $n_0F_{n_0,\nu;\alpha} \le pF_{p,\nu;\alpha}$.

__Remark 2.__ The choice of k associated with the estimator RIDGM, which performed well in the large simulation study of ridge estimators conducted by Dempster, Schatzoff and Wermuth (1977) is equivalent to the solution of $\psi(k) = 1$, where $\boldsymbol{\beta}^* = \boldsymbol{0}$ and $\boldsymbol{T} = \boldsymbol{I}$. For any reasonable α, therefore, the k for RIDGM would be less than k_α.

__Remark 3.__ As a first approximation to k_α, one could approximate $\psi(k)$ by a linear function near $k = 0$; in this neighborhood

$$q^{(k)} \approx (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'k\boldsymbol{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \tag{4.5}$$

so

$$k_\alpha \approx n_0 s^2 F_{n_0,\nu;\alpha}/(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'\boldsymbol{T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*). \tag{4.6}$$

In the case of Hoerl-Kennard ridge regression, where $\boldsymbol{\beta}^* = \boldsymbol{0}$ and $\boldsymbol{T} = \boldsymbol{I}$, (4.6) becomes $n_0 s^2 F_{n_0,\nu;\alpha}/\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$. If additionally we replace $F_{n_2,\nu;\alpha}$ by 1, this becomes the well-known prescription for k proposed by Hoerl, Kennard, and Baldwin (1975).

4.2. Directional compatibility tests.

The compatibility of the external data with the experimental data can also be tested with respect to specific linear combinations $\mu_t = \boldsymbol{t}'\boldsymbol{\beta}$ of interest, where $\boldsymbol{t}'$ is in the row space of $\boldsymbol{X}_0$ (or, equivalently, the row space of $\boldsymbol{T}$). Since

$$\hat{\mu}_t-\mu_t^* = \boldsymbol{t}\hat{\boldsymbol{\beta}}-\boldsymbol{t}'\boldsymbol{\beta}^* \sim N(0,\sigma^2\boldsymbol{t}'((\boldsymbol{X}'\boldsymbol{X})^{-1} + k^{-1}\boldsymbol{T}^-)t), \tag{4.7}$$

a test statistic for compatibility associated with the direction $\boldsymbol{t}$ is:

$$\psi_t(k) = \frac{(\boldsymbol{t}'\hat{\boldsymbol{\beta}} - \boldsymbol{t}'\boldsymbol{\beta}^*)^2}{\boldsymbol{t}'[(\boldsymbol{X}'\boldsymbol{X})^{-1} + k^{-1}\boldsymbol{T}^-]\boldsymbol{t}\, s^2}, \tag{4.8}$$

which is distributed as $F_{1,\nu}$, (i.e., $\psi_t^{1/2}(k)$ is distributed as $|t_\nu|$). Here we have used the notation $\boldsymbol{T}^-$ to refer to the Moore-Penrose generalized inverse of $\boldsymbol{T}$, to cover cases in which $\boldsymbol{T}$ may be singular $(n_0 < p)$.

4.3. Maximum "safe" k.

The direction in which the incompatibility between experimental and external data is most statistically significant can be found by maximizing the numerator of $\psi_t(k)$ in (4.8) while holding the denominator fixed. This is simply a matter of maximizing a non-negative definite quadratic form over a given contour of a positive definite quadratic form. The result is

$$\psi^*(k) = \max_{\underset{\sim}{t}} \psi_t(k) = s^{-2}(\text{max. eigenvalue of } \underset{\sim}{C}(k))$$
$$= s^{-2}q(k) \tag{4.9}$$

where

$$\underset{\sim}{C}(k) = \underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{T})^{-1}k\underset{\sim}{T}(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}^*)(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}^*)', \tag{4.10}$$

and is independent of the particular contour over which the maximization is done. Since the rank of $\underset{\sim}{C}$ is one, the maximum eigenvalue of $\underset{\sim}{C}$ is the same as the trace, which is $q(k)$. The maximizing direction $\underset{\sim}{t}$ is proportional to $\hat{\underset{\sim}{\beta}}-\beta^*_{\sim}$.

We may now define the maximum "safe" k, k'_α, by

$$q(k'_\alpha) = s^2F_{1,\nu;\alpha} \tag{4.11}$$

for some suitably chosen percentage point of the $F_{1,\nu}$ distribution. For $k < k'_\alpha$, no directional compatibility test statistic will be significant at level α.

4.4. Lower bounds for k.

Compatibility tests are useful for setting upper limits on k, but not for setting lower limits, since the closer k is to zero, the more compatible is the external data with the experimental data.

Recall that the reason we resort to the use of external data in the first place is that the data/model for the experiment are inadequate, i.e., the confidence intervals for some estimates of interest are too wide. This point of view immediately establishes as a lower bound for k the smallest value that ensures that the confidence interval (2.18) for all $\underset{\sim}{t}'\underset{\sim}{\beta}$ of interest will be "sufficiently narrow". If this is achieved at $k = 0$, we would be happy with the standard least squares analysis, since we would not need to use the external information, with all its attendant difficulties.

There will be some difficulty in practice if one is unwilling or unable to specify the maximum interval width "necessary" for the interval to be "useful". We would recommend that the end points of the interval (2.18) be plotted as a function of k, or as a function of the more meaningful parameter ρ to be defined in equation (4.13). It should then be apparent how much one gains (in the sense of narrowing the interval) and at what cost (in the sense of increasing the reliance on the external information). This is unavoidably a subjective judgment, but one which we feel is more relevant to the choice of a reasonable value of k than is the "stability" of the Hoerl-Kennard (1970) "ridge trace", a criterion frequently used by practitioners of ridge regression.

Our approach will clearly lead to different choices of k for different end uses, where each end use is the estimation of a linear combination of regression coefficients. For some end uses it may not be necessary to bring in any external information at all, while for others, the external information may need to be weighted heavily. If this is too much trouble, however, one may choose an overall measure of precision to plot against k instead. A reasonable choice would be $\tilde{V}_{avg}(k)$, the average estimated variance of a fitted value over the region of interest in R. Using (2.14) we obtain

$$\tilde{V}_{avg}(k) = \tilde{\sigma}^2(k)\mathrm{tr}\underset{\sim}{M}(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{T})^{-1} \qquad (4.12)$$

where $\underset{\sim}{M}$ is the *region moment matrix* $(E_\phi(\underset{\sim}{x}(d)\underset{\sim}{x}'(d))$ for a uniform probability distribution ϕ over the region of interest) and $\tilde{\sigma}^2(k)$ is given by (2.16). Alternatively, if one were interested only in changes in response from one set of conditions to another, one would replace $\underset{\sim}{M}$ in (4.12) by $2(\underset{\sim}{M} - \underset{\sim}{\xi}\underset{\sim}{\xi}')$ where $\underset{\sim}{\xi} = E_\phi(\underset{\sim}{x}(d))$. Then (4.12) would be the average estimated variance of the difference in predicted response between two randomly chosen points in the region.

4.5. Percent information attributed to external data.

In communicating the results of a regression analysis of the sort we are considering here, it is helpful to provide a quantitative measure of the shares of information attributable to the experimental and external data. Again we rely on the work of Theil (1963), who showed that the function

$$\rho = 100p^{-1}\mathrm{trace}[k\underset{\sim}{T}(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{T})^{-1}] \qquad (4.13)$$

is the only function that meets a certain set of reasonable criteria for expressing the per cent information in $\tilde{\underset{\sim}{\beta}}(k)$ attributable to the external data. Interchanging $\underset{\sim}{X}'\underset{\sim}{X}$ with $k\underset{\sim}{T}$ in (4.13) yields the percent information attributable to the experimental data. This is easily seen to equal $100 - \rho$, as it should. An attractive feature of this definition is that ρ is invariant under nonsingular linear reparametrizations of the model.

In plotting the end points of intervals of type (2.18) as recommended above, it may be useful to use ρ as the horizontal axis rather than k itself. We feel that ρ should be computed in any case, since it is more easily interpretable than k, and since its use places proper emphasis on the fact that $\tilde{\beta}$ is the direct result of incorporating external information.

5. BAYESIAN METHODS

As one might expect, there is a Bayesian parallel to the foregoing regression analysis for fixed k using mixed estimation. The evaluation of the choice of k using tests of compatibility is not a strictly Bayesian concept, however, but is a diagnostic check in the sense of Box (1980).

We shall review here what we would consider to be a standard Bayesian approach, based on "noninformative" priors, and shall indicate that the modifications needed to make it tractable and useful lead to the analysis we have described above. We shall utilize primarily the results of Box and Tiao (1973) and Box (1980). The reader is also referred to the paper of Lindley and Smith (1972) for additional discussion.

We assume the following subjective probability model

$$\begin{aligned} P(\hat{\beta},s^2,\beta^*,\beta,\sigma^2,\sigma_0^2) &\propto (\sigma_0^2)^{-(1+n_0/2)}(\sigma^2)^{-(1+n/2)}(s^2)^{\nu/2-1} \\ &\times \exp[-\frac{1}{2\sigma_0^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - \frac{1}{2\sigma^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \frac{\nu s^2}{2\sigma^2}]. \end{aligned} \tag{5.1}$$

This may be derived by assuming that $\ell n\ \sigma_0^2$, $\ell n\ \sigma^2$, and β^* have independent locally uniform prior distributions and that

$$P(\boldsymbol{\beta}|\boldsymbol{\beta}^*,\sigma^2,\sigma_0^2) \propto (\sigma_0^2)^{-n_0/2} \exp[-\frac{1}{2\sigma_0^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)], \tag{5.2}$$

$$P(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}^*,\boldsymbol{\beta},\sigma^2,\sigma_0^2) \propto (\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})], \tag{5.3}$$

$$P(s^2|\boldsymbol{\beta},\boldsymbol{\beta}^*\boldsymbol{\beta},\sigma^2\sigma_0^2) \propto (s^2)^{\nu/2-1} \exp[-\nu s^2/2\sigma^2]. \tag{5.4}$$

This model is equivalent to that considered by Box and Tiao (1973, Ch. 9), for the estimation of common regresison coefficients using data from two independent sources. Direct adaptation of their results yields the following posterior density for $\boldsymbol{\beta}$:

$$P(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}},s^2,\boldsymbol{\beta}^*) \propto [(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \nu s^2]^{-n/2}[(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)]^{-n_0/2}. \tag{5.5}$$

Unfortunately, this is an improper density since it is not integrable. (The portion of the integral in the neighborhood of $\boldsymbol{\beta}^*$ can be made arbitrarily large.)

An alternative approach would be to fix σ_0^2 by assumption, and then use $P(\hat{\beta},s^2,\beta^*|\sigma_0^2)$ as the basis for a diagnostic check on σ_0^2, in the spirit of Box (1980). The posterior distribution of $\boldsymbol{\beta}$ is then

$$P(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}},s^2,\boldsymbol{\beta}^*,\sigma_0^2) \propto [(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\nu s^2]^{-n/2} \times \exp[-\frac{1}{2\sigma_0^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)]. \qquad (5.6)$$

Although this is a proper density of relatively simple form, we do not know of an easy way to obtain the posterior mean and variance of $\boldsymbol{\beta}$. (One could do it by integrating the moments $E(\boldsymbol{\beta})$ and $E(\boldsymbol{\beta}\boldsymbol{\beta}')$, conditional on $\hat{\boldsymbol{\beta}},s^2,\boldsymbol{\beta}^*,\sigma^2,\sigma_0^2$, numerically with respect to $d\sigma^2/\sigma^2$.) Alternatively, one could seek the mode of (5.6) for use as an estimate of $\boldsymbol{\beta}$. This would require an iterative procedure similar to that suggested by Lindley and Smith (1972), who pointed out, however, that this provides only a point estimate of $\boldsymbol{\beta}$. A quadratic Taylor's series expansion of the logarithm of (5.6) about the mode would lead to an approximate variance-covariance marix. Even if all this were done, however, there are further difficulties in obtaining an appropriate test, in the sense of Box (1980), for use as a diagnostic check on the assumed σ_0^2.

All things considered, therefore, it is not suprising that we should want to fix k by assumption, instead of σ_0^2, as Box (1980) did in an illustration of his estimation /criticism analysis. Box recommended that estimation be based on the (multivariate t) posterior distribution for $\boldsymbol{\beta}$:

$$P(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}},s^2,\boldsymbol{\beta}^*,k) \propto \left[1 + \frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{T})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})}{(\nu+n_0)\tilde{\sigma}^2}\right]^{-(n_0+n)/2} \qquad (5.7)$$

where $\tilde{\boldsymbol{\beta}}$ is given by (2.13) and $\tilde{\sigma}^2$ is given by (2.16). In particular, posterior confidence intervals for any linear combination of the regression coefficients are given by (2.18).

For diagnostic checking, we note that

$$P(\hat{\boldsymbol{\beta}},s^2,\boldsymbol{\beta}^*|k) \propto (s^2)^{\nu/2-1}(q(k)+\nu s^2)^{-(n_0+\nu)/2} \quad . \qquad (5.8)$$

If we regard this as a likelihood function for k, then the likelihood ratio test of any hypothesized k is based on the statistic $\psi(k)$ in (4.1), i.e., it is the same as Theil's test for compatibility between experimental and external information. Box (1980) also derived $\psi(k)$ as one of several diagnostic checks based on the distribution of the experimental data conditional on all the underlying assumptions, and noted its equivalence to Theil's statistic.

REFERENCES

BARNARD, G. A. (1977). On ridge regression, and the general principles of estimation. *Utilitas Mathematics,* 11, 299-311,

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc.,* A-143, 383-404, discussion 404-430.

BOX, G. E. P., and TIAO, G. C.(1973). *Bayesian Inference in Statistical Analysis,* Addison-Wesley, Reading, Mass.

DEMPSTER, A. P., SCHATZOFF, M., AND WERMUTH, N. (1977). A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Assn.,* 72, 77-106.

DRAPER, N. R. AND VAN NOSTRAND, R. C. (1979). Ridge regression and James-Stein estimation: review and comments. *Technometrics,* 21, 451-466.

EGERTON, M. F. AND LAYCOCK, P. J. (1981). Some criticisms of stochastic shrinkage and ridge regression, with counterexamples. *Technometrics,* 23, 155-159.

GENTLEMAN, J. F. AND WILK, M. B. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics,* 31, 387-410.

HOCKING, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

HOERL, A. E., AND KENNARD, R. W. (1970a). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

HOERL, A. E., AND KENNARD, R. W. (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics,* 12, 69-82, Correction 12, 723.

HOERL, A. E., KENNARD, R. W., AND BALDWIN, K. F. (1975). Ridge regression: some simulations. *Comm. in Statist.,* 4, 105-123.

HUBER, PETER J. (1981). *Robust Statistics,* Wiley, New York.

KARSON, M. J., MANSON, A. R. AND HADER, R. J. (1969). Minimum bias estimation and experimental design for response surfaces. *Technometrics,* 11, 461-475.

KUPPER, L. L. AND MEYDRECH, E. F. (1974). Experimental design considerations based on a new approach to a mean square error estimation of response surfaces. *J. Amer. Statistic. Assoc.,* 69, 461-463.

LINDLEY, D. V., AND SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc.,* B-34, 1-41.

MCCABE, G. P., JR. (1978). Evaluation of regression coefficient estimates using α-acceptability. *Technometrics,* 20, 131-139.

OBENCHAIN, R. L. (1977). Classical F-tests and confidence regions for ridge regression. *Technometrics,* 19, 429-439.

SMITH, G. (1980). An example of ridge regression difficulties. *Canadian J. Statist.,* 8, 217-225.

SMITH, G. AND CAMPBELL, F. (1980). A critique of some ridge regression methods. J. Am. Statist. Assoc., 75, 74-81, discussion 81-103.

THIEL, H. (1963). On the use of incomplete prior information in regression analysis. J. Amer. Statist. Assoc., 58, 401-414.

THEIL, H. AND GOLDBERGER, A. S. (1961). On pure and mixed statistical estimation in economics. Int. Econ. Review, 2, 65-78.

THISTED, R. A. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Stanford University Department of Biostatistics Technical Report 28, December.

TJM:NRD/db

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2327	2. GOVT ACCESSION NO. AD-A114 530
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) USING EXTERNAL INFORMATION IN LINEAR REGRESSION, WITH A COMMENTARY ON RIDGE REGRESSION I. MIXED ESTIMATION AND BAYESIAN METHODS	5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Toby J. Mitchell and Norman R. Draper	8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041 DAAG29-80-C-0113
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 4-Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709	12. REPORT DATE January 1982
	13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Collinearity; Dummy data; External information in regression; Ridge regression

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this article, we discuss ways of using "dummy data" and mixed estimation (Theil and Goldberger, 1961), to bring external information formally into linear regression problems when the experimental data/model are inadequate. This is a useful way of attacking the same practical problems that motivated the development of ridge regression (Hoerl and Kennard, 1970).

The main practical issues considered are (i) what form should the "dummy data" take?, and (ii) how much weight should it be given relative to the experimental data? When specific prior information is unavailable, it is —

20. Abstract (continued)

suggested that the dummy data should reflect a preference for "stable" response functions and it is shown how this can be accomplished. Guidelines for the choice of the weighting parameter k (equivalent to the choice of the ridge parameter in ridge regression) are given. Upper limits for k are based on various tests of compatibility between the external (dummy) data and the experimental data. Lower limits for k are determined by the inadequacy of the data/model for the purpose(s) of the analysis. Finally, the parallel Bayesian approach is discussed, with emphasis on Box's (1980) framework of model estimation and criticism.

DATE

FILME

-8